EKONOMETRIA ECONOMETRICS 2(56) • 2017

ISSN 1507-3866
e-ISSN 2449-9994

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

KLASYFIKACJA WIELOMODELOWA DANYCH SYMBOLICZNYCH W BADANIU INNOWACYJNOŚCI KRAJÓW UNII EUROPEJSKIEJ

ENSEMBLE CLUSTERING OF SYMBOLIC DATA IN IDENTIFICATION OF INNOVATION OF EUROPEAN UNION COUNTRIES

DOI: 10.15611/ekt.2017.2.03
JEL Classification: O39; C38; C87

Streszczenie: Innowacje odgrywają coraz to większą rolę w nowoczesnej gospodarce rynkowej. Pozwalają one odnosić korzyści wszystkim obywatelom (producentom, konsumentom i pracownikom). Innowacje mają także kluczowe znaczenie dla poprawy jakości życia, tworzenia lepszych miejsc pracy, a także szeroko rozumianego rozwoju społeczeństwa ekologicznego. Polityka innowacyjności stanowi istotny element polityki na poziomie zarówno krajów, jak i samej Unii Europejskiej. Celem artykułu jest zaprezentowanie przykładu zastosowania podejścia wielomodelowego danych symbolicznych (z zastosowaniem macierzy współwystąpień i metody *k*-medoidów) w klasyfikacji krajów Unii Europejskiej pod względem ich innowacyjności. W części empirycznej wykorzystano pakiety `clusterSim` oraz `symbolicDA` programu R do wykonania obliczeń. W wyniku zastosowania podejścia wielomodelowego zidentyfikowano strukturę czterech różnych klas.

Słowa kluczowe: dane symboliczne, innowacyjność, klasyfikacja wielomodelowa.

Summary: Innovations play a very important part of moden economy. They are the key to higher life quality, better jobs and ecology. The innovation policy is a key element of a national and European Union strategy. The aim of the paper is to present an ensemble clustering of European Union countries considering their innovativeness. This approach allowed to discover four different clusters of countries in innovations. In the empirial part `symbolicDA` and `clusterSim` packages of R software were used. The ensemble approach allowed to obtain four different clusters.

Keywords: symbolic data, innovations, ensemble clustering.

1. Wstęp

Innowacje odgrywają coraz większą rolę w gospodarce rynkowej, dlatego też konieczne jest zrozumienie samego pojęcia innowacji. W literaturze przedmiotu istnieje bardzo wiele różnych ich definicji. Za jedną z pełniejszych można uznać definicję według terminologii OECD, gdzie za działalność innowacyjną uznaje się wiele działań o charakterze naukowym, technicznym, organizacyjnym, handlowym i finansowym, których celem jest opracowanie i wdrożenie nowych lub istotnie ulepszonych produktów i procesów (cyt. za [Górzyński i in. 2004, s. 11]).

Innowacje mają także kluczowe znaczenie dla poprawy jakości życia, tworzenia lepszych miejsc pracy, a także szeroko rozumianego rozwoju społeczeństwa ekologicznego. Polityka innowacyjności stanowi istotny element polityki na poziomie zarówno krajów, jak i samej Unii Europejskiej. Podstawą prawną ogólnej polityki przemysłowej Unii Europejskiej jest art. 173 Traktatu o funkcjonowaniu Unii Europejskiej. Natomiast kwestie badań i rozwoju technologicznego są określone w art. 179-189 Traktatu o funkcjonowaniu Unii Europejskiej. W listopadzie 2014 zatwierdzono wieloletnie środki finansowe na program „Horyzont 2020" na lata 2014-2020. Program ten jest rozwinięciem i kontynuacją wcześniejszych działań podjętych w ramach Strategii lizbońskiej. Zgodnie z przyjętymi założeniami w ramach programu „Horyzont 2020" ma nastąpić ułatwienie wdrożenia w przemyśle innowacyjnych pomysłów naukowych, program ma wspierać interdyscyplinarne, międzysektorowe badania naukowe i innowacje. Program ma także za zadanie wspierać funkcjonowanie i realizację Europejskiej przestrzeni badawczej i Unii innowacji. Wspieranie innowacji ma polegać m.in. na zwiększeniu atrakcyjności zawodu naukowca, zaangażowaniu firm sektora MŚP w badania naukowe i innowacje oraz zwiększeniu uczestnictwa sektora prywatnego w działalności innowacyjnej.

Tematyka innowacyjności, jej pomiaru oraz oceny innowacyjności Polski na tle pozostałych krajów Unii Europejskiej jest szeroko poruszana w literaturze przedmiotu. Warto tu wskazać m.in. prace Stec [2009] i Nowaka [2012], w których dokonano oceny poziomu innowacyjności polskiej gospodarki na tle krajów Unii Europejskiej. Także w pracy Wojtas [2013] dokonano oceny innowacyjności polskiej gospodarki na tle krajów UE. Wykorzystano w niej syntetyczny wskaźnik SII (*Summary Innovation Index*). W pracy Mikołajczyk [2013] dokonano oceny innowacyjności przedsiębiorstw w krajach UE z wykorzystaniem danych z EUROSTAT. Praca Rynardowskiej-Kurzbauer [2015] dokonuje oceny innowacyjności wybranych krajów Europy Środkowo-Wschodniej (w tym Polski) z zastosowaniem syntetycznego indeksu innowacyjności (SII). Omawiane tu prace do oceny innowacyjności zwykle wykorzystują mierniki syntetyczne (zazwyczaj jest to miernik SII), na podstawie których dokonywana jest ocena ogólnego poziomu innowacyjności danego kraju Unii Europejskiej.

Do badania innowacyjności w literaturze przedmiotu zwykle wykorzystuje się mierniki bezpośrednie innowacyjności oraz mierniki pośrednie (mierzące wyniki

działalności wynalazczej), które są oparte na pozytywnym związku pomiędzy poziomem nakładów na badania oraz rozwój a produktywnością i rentownością przedsiębiorstw. Analizując dotychczasowe badania, można zauważyć lukę w zastosowaniu analizy skupień do oceny innowacyjności.

Celem artykułu jest zaprezentowanie przykładu zastosowania podejścia wielomodelowego danych symbolicznych (z zastosowaniem macierzy współwystąpień do łączenia wyników klasyfikacji bazowych i metody *k*-medoidów do otrzymania ostatecznej liczby klas) w klasyfikacji krajów Unii Europejskiej pod względem poziomu ich innowacyjności. W tym celu przeanalizowano wiele indykatorów (wskaźników) innowacyjności dostępnych w Europejskim Urzędzie Statystycznym (EUROSTAT).

2. Podejście wielomodelowe w klasyfikacji danych symbolicznych[1]

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36]:

- zmienne nominalne,
- zmienne porządkowe,
- zmienne przedziałowe,
- zmienne ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli przedziały liczbowe z wagami.

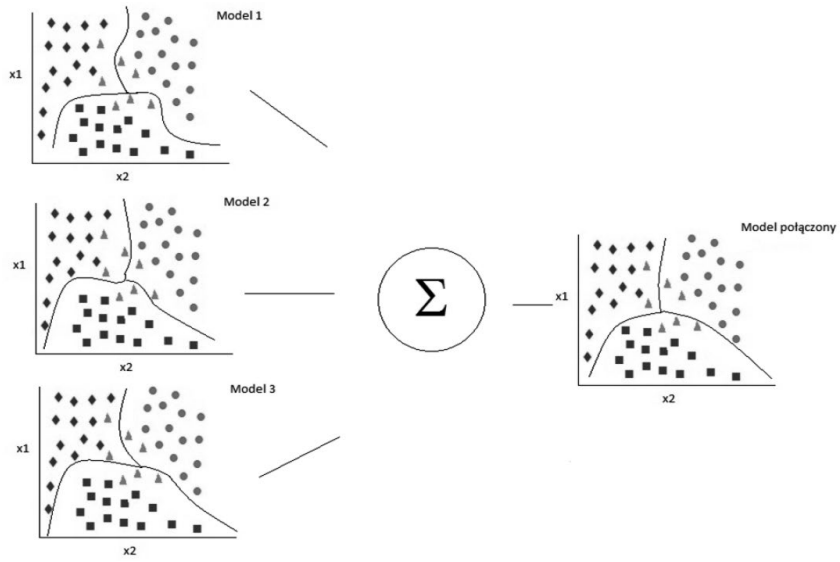
Zastosowanie, oprócz zmiennych klasycznych (nominalnych, porządkowych, przedziałowych, ilorazowych), także zmiennych interwałowych, wielowariantowych, histogramowych czy wielowariantowych z wagami pozwala, z jednej strony, na pełniejszy opis obiektów i zjawisk, ale z drugiej strony, utrudnia analizę danych i wymaga zastosowania odpowiednich metod i narzędzi.

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in.: Bock, Diday (red.) [2000, s. 2-8], Dudek [2013, s. 39-43], Billard, Diday [2006, s. 7-66], Noirhomme-Fraiture, Brito [2011], Diday, Noirhomme-Fraiture [2008, s. 3-30], Gatnar i Walesiak [2011, s. 16--17] i Pełka [2012; 2016].

Podejście wielomodelowe polega na łączeniu wyników otrzymanych za pomocą wielu modeli celem otrzymania jednego, bardziej dokładnego modelu zagregowanego. Idea ta była z powodzeniem stosowana w rozwiązywaniu zagadnień z zakresu dyskryminacji i regresji (zob. np. [Gatnar 2008]) (por. także rys. 1).

---

[1] Szerzej o podejściu wielomodelowym w klasyfikacji danych symbolicznych pisze np. Pełka [2012; 2016].



Rys. 1. Ogólna idea podejścia wielomodelowego w analizie danych

Źródło: opracowanie własne na podstawie pracy Gatnara [2008].

Niemniej idea podejścia wielomodelowego może być z powodzeniem zastosowana także w zagadnieniu klasyfikacji danych symbolicznych. Podejście wielomodelowe w klasyfikacji oznacza łączenie (czyli agregację) wielu klasyfikacji (inaczej modeli) bazowych w jedną klasyfikację złożoną (por. [Fred, Jain 2005]). Celem jest tu otrzymanie bardziej jednorodnej i stabilnej klasyfikacji zagregowanej.

W przypadku podejścia wielomodelowego w analizie skupień dla danych symbolicznych wyróżnia się trzy główne podejścia (por. [Pełka 2012; de Carvalho i in. 2012; Fred, Jain 2005; Pełka 2016; Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999]):

1. Łączenie wyników wielu klasyfikacji bazowych, gdzie zwykle wykorzystuje się macierz współwystąpień do łączenia wyników[2].

2. Łączenie wielu macierzy odległości, z których każda jest traktowana jako odrębny punkt widzenia na zbiór danych. Klasyfikacja polega tu na połączeniu (zagregowaniu) informacji z różnych macierzy odległości[3].

3. Adaptacja metody agregacji bootstrapowej (*bagging*) na potrzeby podejścia wielomodelowego w klasyfikacji[4].

Macierz współwystąpień (*co-association matrix, co-occurence matrix*) jest wynikiem połączenia wielu klasyfikacji bazowych. Różne wyniki klasyfikacji można otrzymać za pomocą zastosowania tej samej metody klasyfikacji, ale z różnymi pa-

[2] Inną metodą jest np. metoda Bordy.

[3] Zob. [de Carvalho i in. 2012].

[4] Zob. [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999].

rametrami, wykorzystania podzbiorów obiektów lub zmiennych w klasyfikacji albo zastosowania różnych metod klasyfikacji (o różnorodnych właściwościach).

Współwystępowanie pary obiektów w tych samych klasach w wielu przeprowadzonych analizach stanowi wskazówkę istnienia związku między nimi. Elementy macierzy współwystąpień są zdefiniowane zgodnie ze wzorem (zob. np. [Fred, Jain 2005, s. 44]):

$$C(i,j) = \frac{n_{ij}}{N}, \tag{1}$$

gdzie: i, j – numer obiektu; n_{ij} – wskazuje, ile razy obiekty i, j znalazły się w tej samej klasie we wszystkich N klasyfikacjach bazowych; N – łączna liczba klasyfikacji bazowych.

Ostateczny podział na klasy otrzymuje się przez wykorzystanie macierzy współwystąpień jako nowej macierzy danych w dowolnej metodzie klasyfikacji (np. hierarchicznej czy iteracyjno-optymalizacyjnej) [Fred, Jain 2005]. Liczbę klas można wyznaczyć za pomocą jednego ze znanych indeksów jakości klasyfikacji albo wykorzystując kryterium najdłuższego wiązania w przypadku klasyfikacji hierarchicznych [Fred, Jain 2005, s. 46-47].

Algorytm klasyfikacji wielomodelowej danych symbolicznych z wykorzystaniem macierzy współwystąpień można przedstawić następująco:

1. Utworzenie S różnych klasyfikacji bazowych na podstawie zbioru danych.
2. Utworzenie na podstawie S klasyfikacji bazowych macierzy współwystąpień zgodnie ze wzorem (1).
3. Zastosowanie macierzy współwystąpień jako nowej macierzy danych w dowolnej metodzie klasyfikacji[5] (w tym artykule wykorzystano metodę k-medoidów).
4. Otrzymanie ostatecznej liczby klas z wykorzystaniem indeksu Bakera-Huberta[6].

3. Klasyfikacja krajów UE pod względem innowacyjności

W badaniu dokonano klasyfikacji 28 krajów Unii Europejskiej na podstawie wskaźników (indykatorów) innowacyjności dostępnych w Europejskim Urzędzie Statystycznym (EUROSTAT).

W celu otrzymania danych symbolicznych interwałowych (w postaci przedziałów liczbowych) dokonano agregacji tych wskaźników w ramach ostatnich pięciu lat (otrzymano w ten sposób tablicę danych symbolicznych). W badaniu uwzględniono następujące zmienne:

---

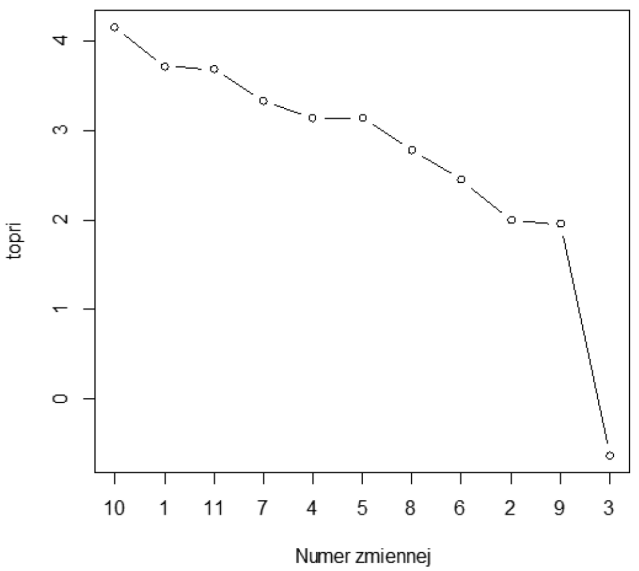
[5] Zob. np. [Fred, Jain 2005].

[6] Ze względu na zastosowanie macierzy współwystąpień jako macierzy danych można zastosować dowolne indeksy jakości klasyfikacji służące wyborowi liczby klas. Zob. np. [Walesiak, Gatnar (red.) 2009, s. 417-418].

- procent PKB przeznaczany na badania w zakresie badań i rozwoju (v_1),
- liczbę osób zatrudnionych w sektorze badań i rozwoju (v_2),
- liczbę wniosków o znak UE na miliard PKB (v_3),
- wielkość eksportu związanego z nowoczesnymi technologiami (w mln euro) (v_4),
- wielkość importu związanego z nowoczesnymi technologiami (w mln euro) (v_5),
- liczbę przedsiębiorstw wysokich technologii (v_6),
- zgłoszenia do Europejskiego Urzędu Patentowego na milion mieszkańców (v_7),
- zgłoszenia do Europejskiego Urzędu Patentowego w ramach biotechnologii na milion mieszkańców (v_8),
- zatrudnionych w sektorze nowoczesnych technologii (jako odsetek zatrudnionych) (v_9),
- zgłoszenia do Amerykańskiego Urzędu Patentowego na milion mieszkańców (v_{10}),
- zgłoszenia do Amerykańskiego Urzędu Patentowego w ramach biotechnologii na milion mieszkańców (v_{11}).

Na podstawie tablicy danych symbolicznych dokonano identyfikacji zmiennych zakłócających z wykorzystaniem metody HINoV (*Heuristic Identification of Noisy Variables*) w wersji zaproponowanej przez Walesiaka i Dudka (zob. [2008]). Metoda ta pozwala zidentyfikować zmienne, które zakłócają istniejącą strukturę klas.

Na rysunku 2 zaprezentowano wykres osypiska dla zmiennych symbolicznych opisujących zbiór danych otrzymany z wykorzystaniem pakietu `symbolicDA` oraz programu R.



Rys. 2. Wartości parametrów *topri* w procedurze HINoV

Źródło: obliczenia własne z wykorzystaniem programu R.

W efekcie zastosowanego algorytmu z pierwotnego zbioru zmiennych zdecydowano się usunąć zmienną v_3 – liczba wniosków o znak UE na miliard PKB.

Dla tak zredukowanej tablicy danych symbolicznych dokonano pomiaru odległości pomiędzy obiektami za pomocą znormalizowanej miary odległości Ichino-Yaguchiego, miary Hausdorffa oraz znormalizowanej miary de Carvalho opartej na potencjale opisowym obiektów symbolicznych[7]. Każdą z miar odległości wykorzystano dla przeprowadzenia pięciu klasyfikacji bazowych z zastosowaniem różnych metod klasyfikacji, co łącznie daje 3 miary odległości × 5 modeli bazowych = 15 klasyfikacji bazowych. Na podstawie klasyfikacji bazowych utworzono macierz współwystąpień, która została wykorzystana w metodzie *k*-medoidów jako macierz danych[8].

Do ustalenia ostatecznej liczby klas wykorzystano indeks Calińskiego-Harabasza (rozważano liczbę klas od 2 do 10) w postaci (zob. np. [Walesiak, Gatnar (red.) 2009, s. 417-418]):

$$\mathrm{G1}(u) = \frac{B_u(u-1)}{W_u(n-u)}, \tag{2}$$

gdzie: u – liczba klas, n – liczba obiektów, $B_u = \mathrm{tr}(\mathbf{B}_u)$, $W_u = \mathrm{tr}(\mathbf{W}_u)$, $\mathbf{B}_u$ – macierz kowariancji międzyklasowej, $\mathbf{W}_u$ – macierz kowariancji wewnątrzklasowej.

Szerzej o indeksach jakości klasyfikacji, które pozwalają wybrać ostateczną liczbę klas, piszą m.in. Walesiak i Gatnar [2009, s. 417-418].

W wyniku przeprowadzonej klasyfikacji wielomodelowej krajów Unii Europejskiej otrzymano strukturę czterech klas (charakterystyki klas ze względu na zmienne symboliczne interwałowe zawarto w tab. 1):

1. Austria, Belgia, Czechy, Finlandia, Francja, Niemcy, Grecja, Irlandia, Włochy, Holandia, Dania. Klasa ta charakteryzuje się największymi różnicami (długościami przedziałów zmiennych symbolicznych) dla wszystkich zmiennych. Są to kraje o dość wysokich, ale zróżnicowanych poziomach innowacyjności.

2. Bułgaria, Chorwacja, Cypr, Estonia, Węgry, Łotwa, Litwa, Luksemburg, Malta, Słowenia. Klasa ta reprezentuje kraje o przeciętnym poziomie innowacyjności. Klasa charakteryzuje się najmniejszą długością przedziałów dla liczby osób zatrudnionych w sektorze B+R, wielkości importu wysokich technologii, liczby przedsiębiorstw wysokich technologii oraz zgłoszeń do Amerykańskiego Urzędu Patentowego w dziedzinie biotechnologii.

3. Portugalia, Rumunia, Słowacja, Hiszpania, Polska. Kraje te charakteryzują się minimalną długością przedziałów dla procenta PKB przeznaczonych na badania

[7] Szerzej na temat miar odległości dla danych symbolicznych piszą m.in. [Dudek 2013, s. 51-61; Gatnar, Walesiak (red.) 2011, s. 18-25].

[8] Do otrzymania ostatecznej liczby klas można zastosować dowolną metodę klasyfikacji. Np. Dudoit i Fridlyanda [2003] stosują metodę *k*-średnich.

i rozwój, wielkości eksportu związanego z eksportem wysokich technologii, zgłoszeń do Europejskiego Urzędu Patentowego (niezależnie od rodzaju patentu), a także Amerykańskiego Urzędu Patentowego w różnych dziedzinach poza biotechnologią. Są to kraje mało innowacyjne.

4. Szwecja, Wielka Brytania. Klasa ta ma najmniejszą długość przedziału dla zmiennej zatrudnionych w sektorze nowoczesnych technologii. Jednocześnie są to kraje wysoce innowacyjne.

Tabela 1. Charakterystyki klas

		Klasy			
		Klasa 1	Klasa 2	Klasa 3	Klasa 4
Zmienne	v_1	<104,2; 1413>	<13,8; 1279>	<15,3; 321,9>	<467,9; 1507,6>
	v_2	<28270; 860842>	<1492; 58237>	<22294; 360229>	<115678; 610276>
	v_4	<715; 176963>	<124; 16861>	<1035; 15250>	<13730; 69322>
	v_5	<2952;147426>	<318; 14036>	<4190; 28491>	<12823; 98357>
	v_6	<0; 144825>	<507; 36679>	<2916; 76741>	<46143; 186761>
	v_7	<6,68; 131>	<0,06; 36,41>	<0,14; 6,21>	<8,78; 91,8>
	v_8	<0,02; 41,8>	<0,03; 5,04>	<0,02; 2,71>	<1,64; 14,73>
	v_9	<28,1; 43,2>	<25,6; 60,4>	<19,2; 32,9>	<41,6; 44,4>
	v_{10}	<1; 99,45>	<0,09; 15,48>	<0,33; 4,25>	<20,14; 83,11>
	v_{11}	<0,03; 19,63>	<0,02; 4,97>	<0,01; 8,95>	<1,93; 8,95>

v_1 – procent PKB przeznaczany na badania w zakresie badań i rozwoju; v_2 – liczba osób zatrudnionych w sektorze badań i rozwoju; v_4 – wielkość eksportu związanego z nowoczesnymi technologiami (w mln euro); v_5 – wielkość importu związanego z nowoczesnymi technologiami (w mln euro); v_6 – liczba przedsiębiorstw wysokich technologii, v_7 – zgłoszenia do Europejskiego Urzędu Patentowego na milion mieszkańców; v_8 – zgłoszenia do Europejskiego Urzędu Patentowego w ramach biotechnologii na milion mieszkańców; v_9 – zatrudnieni w sektorze nowoczesnych technologii (jako odsetek zatrudnionych); v_{10} – zgłoszenia do Amerykańskiego Urzędu Patentowego na milion mieszkańców; v_{11} – zgłoszenia do Amerykańskiego Urzędu Patentowego w ramach biotechnologii na milion mieszkańców.

Źródło: obliczenia własne z wykorzystaniem programu *R*.

4. Podsumowanie

Podejście wielomodelowe może być z powodzeniem zastosowane w klasyfikacji danych symbolicznych różnego typu. W prezentowanym artykule zidentyfikowano strukturę czterech klas o zróżnicowanej strukturze.

Analiza danych symbolicznych pozwala opisywać zbiory danych w pełniejszy sposób niż dane klasyczne. Niemniej jednak wymagają one zastosowania odpowiednich metod klasyfikacji lub zastosowania miar odległości.

Zastosowanie podejścia wielomodelowego pozwoliło zidentyfikować strukturę czterech zróżnicowanych klas. W pierwszej klasie znalazło się jedenaście krajów, które mają dość wysoki, ale mocno zróżnicowany poziom zmiennych charakteryzujących innowacyjność. Poza Republiką Czeską są to kraje tzw. starej Unii.

W klasie drugiej znalazło się dziesięć krajów o przeciętnym poziomie innowacyjności. Są one dość zbliżone do siebie pod względem liczby osób zatrudnionych w sektorze B+R, wielkości importu wysokich technologii, liczby przedsiębiorstw wysokich technologii oraz zgłoszeń do Amerykańskiego Urzędu Patentowego w dziedzinie biotechnologii.

W klasie trzeciej znalazły się kraje o najmniejszym poziomie innowacyjności – Portugalia, Rumunia, Słowacja, Hiszpania, Polska.

Czwarta klasa to Szwecja i Wielka Brytania. Klasa ta ma najmniejszą długość przedziału dla zmiennej zatrudnionych w sektorze nowoczesnych technologii. Jednocześnie są to kraje wysoce innowacyjne.

Literatura

Bock H.-H., Diday E. (red.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg.

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

De Carvalho F.A.T., Lechevallier Y., de Melo F.M., 2012, *Partitioning hard clustering algorithms based on multiple dissimilarity matrices*, Pattern Recognition, 45(1), s. 447-464.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic data analysis. Conceptual statistics and data mining*, Wiley, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Dudek A., Pełka M., Wilk J., 2015, The `symbolicDA` package, www.r-project.org.

Dudoit S., Fridlyand J., 2003, *Bagging to improve the accuracy of a clustering procedure*, Bioinformatics, vol. 19, no. 9, s. 1090-1099.

Fred A.L.N., Jain A.K., 2005, *Combining multiple clustering using evidence accumulation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, s. 835-850.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., Walesiak M. (red.), 2011, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Wyd. C.H. Beck, Warszawa.

Górzyński M., Woodward R., Jakubiak M., 2004, *Innowacyjność polskiej gospodarki w kontekście integracji z UE – możliwości i bariery wdrażania w Polsce gospodarki opartej na wiedzy*, CASE – Centrum Analiz Społeczno-Ekonomicznych, Warszawa.

Hornik K., 2005, *A CLUE for CLUster Ensembles*, Journal of Statistical Software, vol. 14, s. 65-72.

Leisch F., 1999, *Bagged clustering*, Adaptive Information Systems and Modeling in Economics and Management Science, Working Papers, SFB, 51.

Mikołajczyk B., 2013, *Innowacyjność przedsiębiorstw w krajach UE – pomiar i ocena*, Annales Universitatis Mariae Curie-Skłodowska Lublin-Polonia, vol. XLVII, 3, s. 421-431.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, s. 157-170.

Nowak P., 2012, *Poziom innowacyjności polskiej gospodarki na tle krajów UE*, Prace Komisji Geografii Przemysłu, nr 19, s. 153-168.

Pełka M., 2012, *Ensemble approach for clustering of interval-valued symbolic data*, Statistics in Transition, vol. 13, no. 2, s. 335-342.

Pełka M., 2016, *A Comparison Study for Spectral, Ensemble and Spectral Mean-Shift Clustering Approaches for Interval-Valued Symbolic Data*, [w:] Wilhelm A., Kestler H. (red.), *Analysis of Large and Complex Data,* Springer-Verlag, Berlin-Heidelberg, s. 137-146.

Rynardowska-Kurzbauer J., 2015, *Innowacyjność wybranych krajów Europy Środkowo-Wschodniej*, Zeszyty Naukowe Politechniki Śląskiej, seria „Organizacja i Zarządzanie", nr 86, s. 93-101.

Stec M., 2009, *Innowacyjność krajów Unii Europejskiej*, Gospodarka Narodowa, nr 11-12, s. 45-65.

Walesiak M., Dudek A., 2008, *Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis*, [w:] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (red.), *Data Analysis, Machine Learning and Applications*, Springer-Verlag, Berlin-Heidelberg, s. 85-92.

Walesiak M., Dudek A., 2016, The `clusterSim` package, www.r-project.org.

Walesiak M., Gatnar E. (red.), 2009, *Statystyczna analiza danych z wykorzystaniem program R*, PWN, Warszawa.

Wojtas M., 2013, *Innowacyjność polskiej gospodarki na tle krajów Unii Europejskiej*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 756, seria „Finanse, Rynki Finansowe, Ubezpieczenia", nr 57, s. 605-617.